

TREEOME: A framework for epigenetic and transcriptomic data integration to explore regulatory interactions controlling transcription

David M. Budden^{1,2}, Daniel G. Hurley¹ and Edmund J. Crampin^{1,2,3,4,5}

¹Systems Biology Laboratory, Melbourne School of Engineering, ²NICTA Victoria Research Laboratory, ³ARC Centre of Excellence in Convergent Bio-Nano Science and Technology, ⁴Department of Mathematics and Statistics and ⁵School of Medicine, The University of Melbourne, Parkville, Victoria 3010, Australia
{david.budden, daniel.hurley, edmund.crampin}@unimelb.edu.au

Abstract.

Motivation: Predictive modelling of gene expression is a powerful framework for the *in silico* exploration of transcriptional regulatory interactions through the integration of high-throughput -omics data. A major limitation of previous approaches is their inability to handle conditional and synergistic interactions that emerge when collectively analysing genes subject to different regulatory mechanisms. This limitation reduces overall predictive power and thus the reliability of downstream biological inference.

Results: We introduce an analytical modelling framework (TREEOME: tree of models of expression) that integrates epigenetic and transcriptomic data by separating genes into putative regulatory classes. Current predictive modelling approaches have found both DNA methylation and histone modification epigenetic data to provide little or no improvement in accuracy of prediction of transcript abundance despite, for example, distinct anti-correlation between mRNA levels and promoter-localised DNA methylation. To improve on this, in TREEOME we evaluate four possible methods of formulating gene-level DNA methylation metrics, which provide a foundation for identifying gene-level methylation events and subsequent differential analysis, whereas most previous techniques operate at the level of individual CpG dinucleotides. We demonstrate TREEOME by integrating gene-level DNA methylation (bisulfite-seq) and histone modification (ChIP-seq) data to accurately predict genome-wide mRNA transcript abundance (RNA-seq) for H1-hESC and GM12878 cell lines.

Availability: TREEOME is implemented using open-source software and made available as a pre-configured bootable reference environment. All scripts and data presented in this study are available online at <http://sourceforge.net/projects/budden2015treeome/>.

Contact: edmund.crampin@unimelb.edu.au

1 Introduction

Understanding the precise spatiotemporal regulation of eukaryotic gene expression is a central challenge in molecular biology. Transcriptional regulation is governed by dynamic restructuring of chromatin to control gene accessibility, mediated by post-translational modifications of nucleosomal histone proteins. Any perturbation of the systems regulating gene accessibility can affect critical cellular functions including homeostasis, differentiation and apoptosis. Consequently, dysregulation of these systems has been implicated with hundreds of developmental, autoimmune, neurological, inflammatory and neoplastic disorders [1].

The relationship between histone modifications and gene expression involves complex systems of protein-mediated regulatory events that are still poorly understood. The simplest interactions involve acetylation of lysine residues on the histone H3/4 amino-termini, reducing their net-positive charge and weakening charge-dependent interactions with adjacent nucleosomes and the negatively-charged DNA backbone [2]. Promoter-localised histone acetylation is thus considered a euchromatic modification, as it promotes the establishment of

open DNase-sensitive chromatin and active transcription. Histone lysine methylation is further separated from the physical transcription process; mono-, di- and tri-methylation of specific residues are recognised by proteins with varying and context-sensitive regulatory roles, including the Polycomb repressive complexes (associated with H3K27me2/3) and DNA *de novo* methyltransferase family (associated with H3K9me2/3).

To develop a comprehensive understanding of the regulatory logic controlling eukaryotic gene expression by studying individual protein-protein interactions would require a currently-unavailable volume and resolution of proteomics data. Instead, predictive modelling frameworks have been developed that leverage the wealth of high-throughput sequencing data generated by recent large-scale consortia (*e.g.* [3]) to study the indirect relationships between histone modifications and transcript abundance. The utility of these models is not only the ability to predict RNA abundance for individual species (at which the best models currently available perform rather poorly at the level of individual genes), but rather the biological insights that can be gained by exploring the relationships inferred from the data; the prediction accuracy of such models is simply an indirect measure of their explorative potential.

We have previously reviewed predictive modelling in the context of eukaryotic transcriptional regulation [4], which has been applied to a wide range of problems in molecular biology. These include: inferring regulatory roles of transcription factors from their respective binding motifs [5]; identifying regulatory elements responsible for differential expression patterns [6]; exploring the relationship between gene expression and higher-order chromatin domains [7]; and large-scale comparative analysis of the transcriptome across distant species [8].

Despite the utility of predictive modelling as a framework for uncovering novel molecular biology, a major limitation of current approaches is their inability to handle conditional and synergistic interactions that emerge when analysing genes subject to different regulatory mechanisms. For this study, we have selected four histone modifications that epitomise this problem through their indirect and context-sensitive regulatory roles: H3K4me3, H3K27me3, H3K9me3 and the H2A.Z histone variant. A simplified histone/epigenetic code for these modifications is illustrated in Figure 1.

As an example of regulatory heterogeneity, Figure 1 illustrates that H3K4me3 (commonly associated with gene activation [2]) promotes transcription in the absence of H2A.Z, but in the presence of H2A.Z only promotes transcription if H3K27me3 is absent [9]. This description of conditional behaviour remains an oversimplification of the underlying regulatory events, as the transcriptional effect of histone lysine methylation depends on both the level of methylation (mono/di/tri), location within the gene (5' versus 3') and the presence of other regulatory elements (*e.g.* transcription factors and ncRNAs). Current predictive modelling approaches generally assume linear/additive models and are unable to capture these conditional relationships. Although some studies have investigated the application of non-linear regression models (*e.g.* support vector regression [10]), quantitative analysis of these models across multiple scenarios has revealed that they perform no better than standard linear models [4]. We speculate that this analytical limitation is a major cause of the distinct lack of models integrating DNA methylation data in previous studies, as low-expression genes may be under the control of a variety of other silencing mechanisms.

In this study, we introduce an analytical framework (TREEOME: tree of models of expression) that facilitates the integration of regulatory data by separating genes into putative regulatory classes on the basis of histone modification and/or DNA methylation state. We demonstrate TREEOME by integrating gene-level DNA methylation (bisulfite-seq) and histone modification (ChIP-seq) data to predict genome-wide RNA transcript abundance (RNA-seq) for H1-hESC and GM12878 cell lines. As there has been little previous work in formulating and/or evaluating gene-level DNA methylation statistics, our analysis is prefaced by a quantitative evaluation of four possible promoter-localised methylation scores. These methylation scores provide a foundation for identifying significant methylation and subsequent differential analysis, at the level of genes rather than individual CpG dinucleotides.

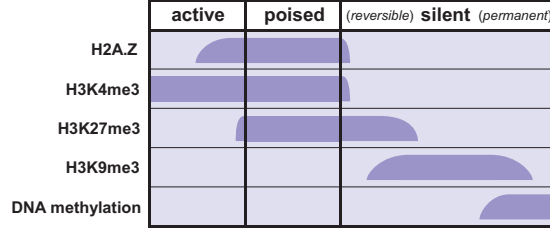


Fig. 1. Illustration of the histone/epigenetic code in the context of the promoter-localised regulatory elements analysed in this study. Only active genes exhibit significant expression, corresponding with H3K4me3 often flanked by H2A.Z. Poised and reversible/permanently silenced genes are distinguished by decreasing likelihood of genes returning to an active state; poised genes are marked by bivalent H3K4/27me3 and H2A.Z, while silent genes are marked by H3K27me3 (facultative heterochromatin), H3K9me3 (constitutive heterochromatin) and DNA methylation (permanent silencing).

2 Methods and materials

2.1 Gene-specific histone scores

The association strength between each gene, i , and histone modification, j , is calculated using the constrained sum-of-tags histone score [4]:

$$a_{ij} = \sum_k g_k, \quad (1)$$

where g_k is the number of ChIP-seq reads (or normalised equivalent) for j mapped to position k relative to the TSS of i . As ChIP-seq involves sequencing of DNA corresponding with the end of each nucleosome, the position for each read was shifted by ± 73 bp (for \pm strand respectively) to centre on the modified nucleosome [11]. Integrating over a region 2000 bp either side of the TSS (approximating the average width of histone modification ChIP-seq binding regions) is standard for this approach [10,6,5] and applied throughout this study.

2.2 Predictive modelling of gene expression

In this study, we model the RPKM-normalised transcript abundance, y_i , of each gene, i , as a general linear function of its association, a_{ij} , with each histone modification, j :

$$\sinh^{-1}(y_i) = \mu + \sum_j \beta_j a_{ij} + \varepsilon_i, \quad (2)$$

where β_j captures the influence of histone modification j on gene expression, μ is the basal expression level, and ε_i is the gene-specific error term. The inverse hyperbolic sine (arsinh) transformation, $\sinh^{-1}(x) = \log(x + \sqrt{1 + x^2})$, is approximately equal to $\log(2x)$ for $x \gg 0$, allowing it to be regarded as practically-equivalent to the log-transformation applied in previous gene expression modelling studies [4]. Unlike $\log(x)$, $\sinh^{-1}(x)$ is defined for $x = 0$, removing the need to meta-optimize small constants to add to x (leading to spurious inflation of prediction accuracy) and making it better-suited to integrating ChIP-seq and RPKM-normalised RNA-seq data.

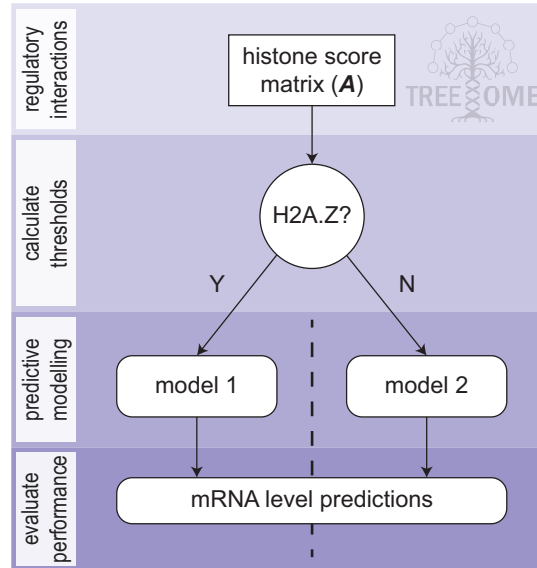


Fig. 2. Illustration of a general predictive modelling pipeline where the H2A.Z histone variant has been used to separate genes into two categories. Categorising genes by the presence of promoter-localised H2A.Z removes significant heterogeneity in the regulatory role of H3K4me3; H3K4me3 in the presence of H2A.Z is often a hallmark of low expression (*i.e.* poised genes), whereas H3K4me3 is otherwise associated with active transcription. These synergistic interactions are poorly-handled by current regression modelling.

2.3 Modelling conditional regulatory interactions with decision trees

The decision tree framework (TREEOME) mitigates the analytical consequences of conditional and synergistic interactions in gene expression data. For example, gene-level H2A.Z scores (an indicator of histone bivalency) could be used to separate genes into two subsets: those putatively-regulated by H2A.Z and those that are not. Separate predictive models can then be constructed and evaluated for both subsets from the remaining regulatory elements, as illustrated in Figure 2.

TREEOME uses an unsupervised method to define the threshold above which a gene-level histone score is accepted to represent actual regulatory activity:

- Under the assumption that H2A.Z is sufficient to separate genes into two sets of homogeneously-regulated genes (or where both sets are to be further subdivided according to other epigenetic markers), the threshold is chosen to maximise the combined prediction accuracy of both models.
- Under the assumption that only one subset is homogeneously-regulated (*i.e.* the other is to be further subdivided), the threshold is chosen to maximise the prediction accuracy of the homogenous model.

TREEOME implements a greedy algorithm that is not necessarily globally optimal. Although improved prediction accuracy could be obtained by optimising over the full set of thresholds for an arbitrarily-large tree, this approach would lose the biological meaning (regulated-or-not) underlying our threshold selection methodology.

2.4 Evaluation of prediction accuracy

Prediction accuracy is assessed for each regression model using an adjusted R^2 score, which in comparison to the standard R^2 approach prevents spurious inflation of the statistic due to introduction of additional explanatory variables [12].

Separate RNA-seq replicates were used for model training and evaluation to prevent over-fitting to experimental noise. If multiple replicates are not available, the adjusted R^2 score for each model can be determined using a k -fold cross-validation process.

2.5 Derivation of putative regulatory roles

Putative regulatory roles are inferred for each histone modification using principal component analysis (PCA). Specifically, the histone score matrix, \mathbf{A} (see eq. (1)), for a gene-set of interest is arsinh-transformed and reformulated using the following singular value decomposition [13]:

$$\sinh^{-1}(\mathbf{A}) = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \quad (3)$$

where \mathbf{U} is the matrix of component scores, $\mathbf{\Sigma}$ is the diagonal matrix of the singular values of \mathbf{A} , and \mathbf{V} is the matrix of loadings (weights by which the histone scores are multiplied to derive their respective component scores). In the context of modelling gene expression, the columns of the matrix $\mathbf{U}\mathbf{\Sigma}$ are the principal components (PCs), and the rows correspond with eigengenes [14]. The data-derived, putative regulatory role of each histone modification is simply its contribution (loading) toward the individual PC most predictive of gene expression [4].

2.6 Quantifying gene-level DNA methylation

Compared to CpG-level methylation scores, gene/region-level DNA methylation scores are not well-established in previous literature. We explore four possible promoter-localised scores in the context of predictive gene expression modelling, considering a window 2000 bp either side of the respective gene's TSS:

- **Sum of methylation fractions by site (SMFS)**: Sum of the CpG-level methylation scores within a region, similar to the constrained sum-of-tags score previously applied to the analysis of ChIP-seq data [4]
- **Mean methylation fraction by site (MMFS)**: Equivalent to the SMFS score divided by the number of assayed CpGs within the region, similar to the mean methylation level described by [15]
- **Mean methylation fraction by region (MMFR)**: Proportion of raw reads that were found to be methylated, similar to the weighted methylation level described by [15]
- **Sum of scaled methylation reads by region (SMRR)**: Equivalent to the MMFR score where each read is multiplied by $-\exp(d/d0)$, where d is the distance (bp) from the TSS and $d0 = 5000$, similar to the exponentially decaying affinity score previously applied to the analysis of ChIP-seq data [4]

2.7 Data

H1-hESC and GM12878 data were selected to demonstrate TREEOME for both pluripotent and differentiated cell lines, as functional patterns of DNA methylation vary significantly across the lineage commitment spectra. All H1-hESC and GM12878 gene expression (RNA-seq), histone modification (ChIP-seq) and DNA methylation (methyl RRBS) data were downloaded from ENCODE [3]. Specific GEO accession numbers for each dataset are provided in Table 1. The TSS for each gene was taken from the gene annotation dataset for the human genome (hg19/GRCh37). Multiple transcripts or isoforms were removed by considering only the most 5'-located TSS for each unique Ensembl gene identifier, resulting in a set of 11,806 genes for analysis. RNA-seq data was re-mapped to hg19 using Subread [16] and RPKM-normalised using edgeR [17,18].

Table 1. All H1-hESC and GM12878 data used in this study [3].

Data type	Data source
RNA-seq	GSM958730 (GM12878, 2 replicates) GSM958737 (H1-hESC, 2 replicates)
TSS	Ensembl hg19/GRCh37 [19]
Methyl RRBS (GM12878)	GSM683906 (replicate 1) GSM683927 (replicate 2)
ChIP-seq (GM12878)	GSM733767 (H2A.Z) GSM733758 (H3K27me3) GSM733708 (H3K4me3) GSM733664 (H3K9me3)
Methyl RRBS (H1-hESC)	GSM683770 (replicate 1) GSM683879 (replicate 2)
ChIP-seq (H1-hESC)	GSM1003579 (H2A.Z) GSM733748 (H3K27me3) GSM733657 (H3K4me3) GSM1003585 (H3K9me3)

2.8 Implementation

TREEOME is implemented using open-source software and made available as a pre-configured bootable virtual environment using the approach described by [20]. This environment was created using a minimal installation of Ubuntu 13.10; a lightweight Linux distribution which supports all the tools required. R version 3.0.1 was installed, along with the core set of packages and utilities required to explore the presented results. Alternatively, all data and scripts are available online at

<http://sourceforge.net/projects/budden2015treeome/>.

3 Results and discussion

We used TREEOME to build models based on dividing H1-hESC and GM12878 data into gene sets for regression modelling on the basis of DNA methylation score and subsequently (for genes identified as not methylated) on the basis of H2A.Z, a well-studied indicator of histone bivalency and poised expression [9]. First, we establish which combination of gene-level histone and DNA methylation scores are most appropriate for our subsequent analyses, as described below.

3.1 Histone modifications are predictive of transcript abundance

To validate whether our histone score (eq. 1) and regression model (eq. 2) formulations are suitable for the data considered in this study, we evaluated the accuracy of model-predicted RPKM-normalised transcript abundance compared to actual RNA-seq data genome-wide for each cell line. These results are presented in Figure 3, and the performance of our models (adjusted $R^2 = 0.43$ for H1-hESC and 0.47 for GM12878) were found to be similar to those of previous studies [10,6,11,5].

Figure 3 also presents the distribution of gene expression levels and data-derived putative regulatory roles of each histone modification genome-wide (methodology described in Sec. 2.5), with positive/negative loadings suggesting activator/repressor roles respectively. It is evident that the differentiated lymphoblastoid GM12878 cell line exhibits more near-zero expression (silenced) genes than pluripotent H1-hESC, as expected due to DNA methylation-mediated gene silencing during lineage commitment. DNA methylation is further implicated by the stronger regulatory signal for H3K9me3 in GM12878, which is associated with DNA *de novo* methyltransferase activity [21].

3.2 MMFS is the most informative methylation score

It is widely accepted that promoter-localised CpG methylation prevents the initiation of eukaryotic gene transcription [22]. By extension, a suitable gene-level DNA methylation score should be anti-correlated with transcript abundance derived from genome-wide RNA-seq data. Figure 4 presents the correlation between transcript abundance and each of the four DNA methylation scores described in Section 2.6 (SMFS, MMFS, MMFR and SMRR) for all replicate combinations. MMFS performed equal-best for H1-hESC (Pearson’s $r = -0.25$) and outright best for GM12878 (Pearson’s $r = -0.31$), with all scores exhibiting stronger anti-correlation with GM12878 than hESC, as expected from Section 3.1.

The distribution of promoter methylation (MMFS) versus transcript abundance presented in Figure 4 demonstrates two distinct clusters, corresponding with active/unmethylated (green) and silenced/methylated genes (red). It is also evident that a large number of genes exhibit near-zero expression despite a lack of substantial DNA methylation (blue); these genes reduce the predictive power of DNA methylation in a current modelling framework and are likely silenced by other mechanisms (*e.g.* repressor/silencer transcription factors [23] or H3K27me3-mediated Polycomb activity [21]).

3.3 Naïve predictive model integration is unsuitable for DNA methylation data

As demonstrated in Figure 4, all four gene-level DNA methylation scores are anti-correlated with genome-wide RNA transcript abundance, as expected due to the well-established silencing role of promoter-localised CpG methylation [22]. Intuitively, integrating any of these scores into a gene expression model (particularly MMFS) should yield improved prediction accuracy due to the addition of information regarding methylation-mediated silencing.

A naïve approach to integrating DNA methylation into the predictive modelling framework (described in Section 2.2) involves simply concatenating the vector of methylation scores as a new column of the $n \times m$ histone score matrix, \mathbf{A} , where n is the number of genes and m is the number of histone modifications. We constructed these models for all combinations of cell line and DNA methylation score and found that the resultant change in prediction accuracy was negligible in all cases ($|\Delta \text{adj.} R^2| < 10^{-3}$).

Strikingly, despite the anti-correlation shown between each methylation score and RNA transcript abundance, the naïve integration of this information into predictive models trained on histone modification data yields practically-zero improvement in prediction accuracy (irrespective of score or cell line). Within the constraints of a linear regression framework, DNA methylation and the four considered histone modifications are statistically redundant with respect to gene expression (similar redundancy between histone modifications and transcription factors has recently been explored in detail by [7]). These results indicate that a more principled approach of integrating transcriptional regulatory data is necessary to better leverage biological insight from predictive models.

3.4 TREEOME improves predictive power for homogeneous regulatory classes

We use the best-performing methylation score (MMFS) to separate genes into two putative regulatory classes (MMFS⁺ versus MMFS[−]). Intuitively, this approach should isolate genes subject to H3K9me3/DNA methylation-mediated silencing from an otherwise-heterogeneous set.

Table 2. Proportion of genes attributed to each putative regulatory class and respective improvement in prediction accuracy $\Delta\text{adj.}R^2$ (relative to a traditional model constructed from the same data) for both H1-hESC and GM12878 cell lines. Adjusted R^2 scores were calculated using separate RNA-seq replicates for training and evaluation, as described in Section 2.4.

	MMFS ⁺		H2A.Z ⁺		H2A.Z ⁻	
	genes	$\Delta\text{adj.}R^2$	genes	$\Delta\text{adj.}R^2$	genes	$\Delta\text{adj.}R^2$
H1-hESC	46%	+0.03	25%	-0.01	28%	+0.05
GM12878	40%	+0.06	29%	-0.13	30%	+0.16

Unmethylated genes are still subject to a variety of transcriptional regulatory mechanisms, including H3K4me3-mediated euchromatinisation (activation) and H3K27me3-mediated facultative heterochromatinisation (repression) [24]. As described in Section 2.3, our ability to identify the signatures of these mechanisms is confounded by bivalency, where the otherwise antagonistic H3K4/27me3 are maintained in metastable equilibrium by the H2A.Z histone variant [9]. Therefore, to further remove synergistic effects from our predictive models, the aforementioned set of MMFS⁻ genes is separated into two further putative regulatory classes by H2A.Z score (H2A.Z⁺ and H2A.Z⁻). The final decision tree structure is illustrated in Figure 5.

In addition to the decision tree structure, Figure 5 demonstrates the following for the H1-hESC cell line: the threshold selection process (described in Section 2.3); the proportion of genes attributed to each putative regulatory class; and the respective performance results ($\Delta\text{adj.}R^2$) relative to an unseparated regression model constructed from the same data. The statistics for both H1-hESC and GM12878 TREEOME analyses are presented in Table 2.

By separating genes into subsets exhibiting greater regulatory homogeneity, it is evident from Table 2 that the inferred relationships between regulatory and expression data are significantly strengthened for the majority of genes; *e.g.* 40% of GM12878 genes are classified as MMFS⁺, and our ability to predict the expression of these genes improves significantly ($\Delta\text{adj.}R^2 = 0.06$, yielding a model with an overall predicted-versus-measured transcript abundance correlation of Pearson’s $r > 0.70$).

Reduction in prediction accuracy is constrained to H2A.Z⁺ genes, which we speculate is due to inherent heterogeneity in H2A.Z-mediated regulation; *i.e.* H2A.Z is known to both maintain H3K4/27me3 bivalency and flank the TSS during transcriptional activation [26]. It is likely that integrating further histone modifications or related data (*e.g.* DNase-I hypersensitivity) would allow TREEOME to resolve this heterogeneity.

4 Conclusions

In this study, we have demonstrated that a decision tree-based analytical framework (TREEOME) is able to improve prediction accuracy of regression models for genome-wide RNA transcript abundance by separating genes into putative regulatory classes. We demonstrated the effectiveness of TREEOME by providing the first integration of DNA methylation (bisulfite-seq) and histone modification (ChIP-seq) data to accurately predict genome-wide RNA transcript abundance (RNA-seq) for H1-hESC and GM12878 cell lines.

As described in Section 1, the utility of predictive gene expression modelling is not the ability to predict RNA levels, but rather the insights into epigenetic regulation of gene expression that can be gained by exploring the relationships inferred from the data. Figure 6 illustrates one of many possible examples of a predictive modelling workflow, in the context of inferring the unknown regulatory roles of a transcription factor from its position weight matrix.

One limitation of this approach is that it assumes that the functional role of a bound transcription factor is independent of the local chromatin landscape. This is certainly not the case; *e.g.* pioneer transcription factors are able to directly engage nucleosomal DNA [28], although translating their activating effect to proximal

genes requires a subsequent cascade of chromatin-remodelling events that is impossible in the presence of DNA methylation [22]. This limitation is removed by integrating DNA methylation data (via TREEOME) into the blue-highlighted component of Figure 6 (as described in Section 2.3), allowing the derivation of separate regulatory roles and associated confidence values (prediction accuracy) for both methylated and unmethylated genes.

Our TREEOME analysis was prefaced by the first quantitative evaluation of several methods of quantifying gene-level DNA methylation events, which have widespread potential in facilitating future gene-level (rather than CpG-level) differential methylation analyses. We found that the (promoter-localised) mean methylation fraction by site (MMFS) score yields the greatest anti-correlation with gene expression levels in both H1-hESC and GM12878.

We have endeavoured to demonstrate the utility of TREEOME in a practical context and unsupervised manner. The four histone modifications were selected due to their highly conditional and indirect regulatory influence; integrating elements with more direct effects (*e.g.* histone lysine acetylations or DNase I hypersensitivity) would undoubtedly improve prediction accuracy, as demonstrated in previous studies (*e.g.* [11]). The unsupervised TREEOME threshold selection process could likewise be replaced to capture prior biological knowledge (*e.g.* known methylated genes) or directly-optimised against global prediction accuracy, although we maintain that the latter approach would lose the biological meaning (regulated-or-not) underlying our methodology.

Acknowledgement

This work was supported by an Australian Postgraduate Award [D.M.B.]; the Australian Federal and Victoria State Governments and the Australian Research Council through the ICT Centre of Excellence program, National ICT Australia (NICTA) [D.M.B., E.J.C.]; and the Australian Research Council Centre of Excellence in Convergent Bio-Nano Science and Technology (project number CE140100036) [E.J.C.]. The views expressed herein are those of the authors and are not necessarily those of NICTA or the Australian Research Council.

References

1. Portela, A., Esteller, M.: Epigenetic modifications and human disease. *Nature Biotechnology* **28**(10) (2010) 1057–1068
2. Bannister, A.J., Kouzarides, T.: Regulation of chromatin by histone modifications. *Cell Research* **21**(3) (2011) 381–395
3. ENCODE Project Consortium: An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414) (2012) 57–74
4. Budden, D.M., Hurley, D.G., Crampin, E.J.: Predictive modelling of gene expression from transcriptional regulatory elements. *Briefings in Bioinformatics* doi: **10.1093/bib/bbu034** (2014)
5. McLeay, R.C., Lesluyes, T., Partida, G.C., Bailey, T.L.: Genome-wide *in silico* prediction of gene expression. *Bioinformatics* **28**(21) (2012) 2789–2796
6. Cheng, C., Gerstein, M.: Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic Acids Research* **40**(2) (2012) 553–568
7. Budden, D.M., Hurley, D.G., Cursons, J., Markham, J.F., Davis, M.J., Crampin, E.J.: Predicting expression: The complementary power of histone modifications and transcription factor binding data. *Epigenetics and Chromatin* (2014)
8. Gerstein, M.B., Rozowsky, J., Yan, K.K., Wang, D., Cheng, C., Brown, J.B., Davis, C.A., Hillier, L., Sisu, C., Li, J.J., et al.: Comparative analysis of the transcriptome across distant species. *Nature* **512**(7515) (2014) 445–448
9. Voigt, P., Tee, W.W., Reinberg, D.: A double take on bivalent promoters. *Genes & Development* **27**(12) (2013) 1318–1338
10. Cheng, C., Yan, K.K., Yip, K.Y., et al.: A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biology* **12**(2) (2011) R15
11. Karlić, R., Chung, H.R., Lasserre, J., et al.: Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences* **107**(7) (2010) 2926–2931

12. Harel, O.: The estimation of r^2 and adjusted r^2 in incomplete data sets using multiple imputation. *Journal of Applied Statistics* **36**(10) (2009) 1109–1118
13. Watkins, D.S.: *Fundamentals of matrix computations*. Volume 2. John Wiley & Sons (2004)
14. Alter, O., Brown, P.O., Botstein, D.: Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences* **97**(18) (2000) 10101–10106
15. Schultz, M.D., Schmitz, R.J., Ecker, J.R.: Leveling the playing field for analyses of single-base resolution DNA methylomes. *Trends in Genetics* **28**(12) (2012) 583
16. Liao, Y., Smyth, G.K., Shi, W.: The subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research* **41**(10) (2013) e108–e108
17. Mortazavi, A., Williams, B.A., McCue, K., et al.: Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**(7) (2008) 621–628
18. Robinson, M.D., McCarthy, D.J., Smyth, G.K.: edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**(1) (2010) 139–140
19. Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al.: Ensembl 2014. *Nucleic Acids Research* (2013) gkt1196
20. Hurley, D.G., Budden, D.M., Crampin, E.J.: Virtual reference environments: A simple way to make research reproducible. *Briefings in Bioinformatics* **doi: 10.1093/bib/bbu043** (2014)
21. Cedar, H., Bergman, Y.: Linking DNA methylation and histone modification: patterns and paradigms. *Nature Reviews Genetics* **10**(5) (2009) 295–304
22. Jones, P.A.: Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics* **13**(7) (2012) 484–492
23. Spitz, F., Furlong, E.E.: Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics* **13**(9) (2012) 613–626
24. Li, B., Carey, M., Workman, J.L.: The role of chromatin during transcription. *Cell* **128**(4) (2007) 707–719
25. Zilberman, D., Coleman-Derr, D., Ballinger, T., Henikoff, S.: Histone H2A. Z and DNA methylation are mutually antagonistic chromatin marks. *Nature* **456**(7218) (2008) 125–129
26. Raisner, R.M., Hartley, P.D., Meneghini, M.D., Bao, M.Z., Liu, C.L., Schreiber, S.L., Rando, O.J., Madhani, H.D.: Histone variant H2A. Z marks the 5 ends of both active and inactive genes in euchromatin. *Cell* **123**(2) (2005) 233–248
27. Cuellar-Partida, G., Buske, F.A., McLeay, R.C., Whittington, T., Noble, W.S., Bailey, T.L.: Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics* **28**(1) (2012) 56–62
28. Zaret, K.S., Carroll, J.S.: Pioneer transcription factors: establishing competence for gene expression. *Genes & Development* **25**(21) (2011) 2227–2241

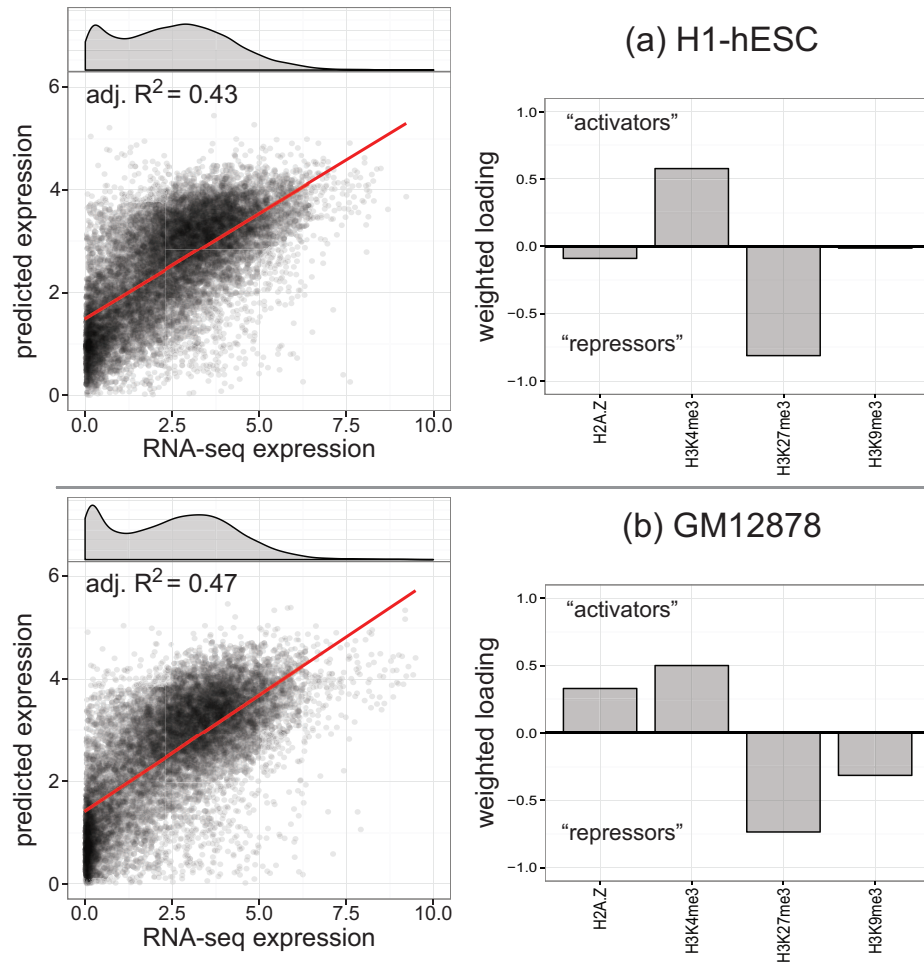


Fig. 3. Analysis of predictive models of genome-wide transcript abundance for (a) H1-hESC and (b) GM12878 cell lines, constructed from H2A.Z, H3K4me3, H3K27me3 and H3K9me3 histone scores. Both panels demonstrate the following: (top-left) the distribution of arsinh-transformed RPKM-normalised transcript abundance derived from RNA-seq data; (left) predicted-versus-measured transcript abundance for the linear regression model, with performance quantified as an adjusted R^2 score; and (right) the data-derived putative regulatory roles of each histone modification, with positive/negative loadings suggesting activator/repressor roles respectively. Of particular interest is the latent signature of DNA methylation-mediated gene silencing, with GM12878 exhibiting a higher proportion of near-zero expression genes and strikingly stronger regulatory signal for H3K9me3 (implicated in DNA *de novo* methyltransferase activity), as expected following lineage-commitment.

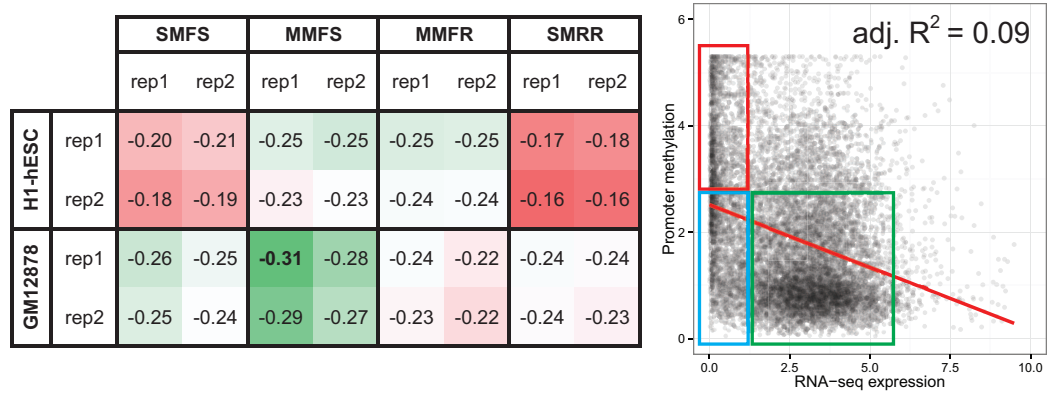


Fig. 4. Analysis of the gene-level DNA methylation scores described in Section 2.6 (SMFS, MMFS, MMFR and SMRR). (Left) MMFS exhibits the strongest overall anti-correlation with RPKM-normalised transcript abundance (Pearson's $r = -0.31$), indicating that it is most appropriate for capturing the gene silencing effect of promoter-localised methylation. Model performance colour-coded by correlation, with the best/worst-performing models highlighted in green/red respectively. (Right) promoter methylation (MMFS) versus transcript abundance genome-wide for GM12878 (regression line shown in red), demonstrating two distinct gene clusters: active/unmethylated (green) and silent/methylated (red). It is also evident that a large number of genes exhibit near-zero expression despite a lack of substantial DNA methylation (blue); these genes reduce the predictive power of DNA methylation genome-wide and are likely silenced by other mechanisms (*e.g.* repressor/silencer transcription factors [23] or H3K27me3-mediated Polycomb activity [21]).

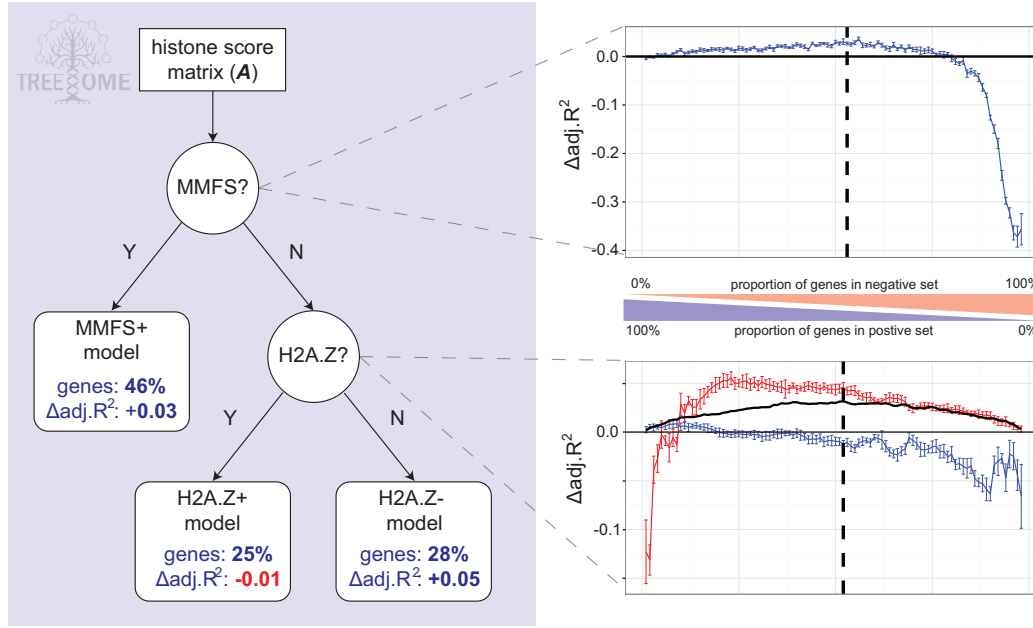


Fig. 5. Decision tree of predictive models of H1-hESC RNA transcript abundance, constructed from the same data as the model described in Section 3.3. This tree uses promoter-localised DNA methylation (MMFS) and the H2A.Z histone variant to classify genes into three putative regulatory classes: MMFS^+ (high MMFS score), H2A.Z^+ (low MMFS and high H2A.Z) and H2A.Z^- (low MMFS and low H2A.Z). A fourth category (high/high) would be biologically meaningless as DNA methylation and H2A.Z are mutually exclusive *in vivo* [25]. Thresholds were learned directly from the data using the unsupervised approach described in Section 2.3. Specifically, the blue, red and black lines illustrate $\Delta\text{adj.}R^2$ (relative to a standard model constructed from the same data) as a function of threshold values for positive (e.g. high MMFS), negative and cumulative models respectively, with the optimal value for both forks indicated by a black dashed line. Error bars capture the standard error of the mean ($\mu \approx 0$) for models constructed from 100 randomly-sampled gene-sets of equal size, illustrating the performance variation expected by chance (i.e. fewer genes equals larger variation in model performance, as expected).

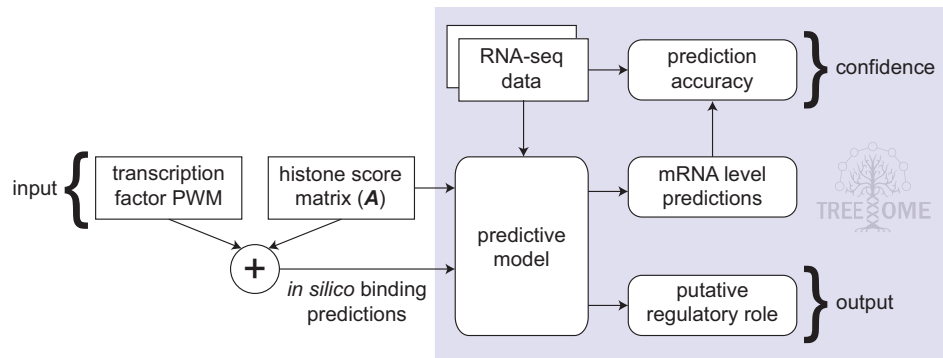


Fig. 6. Example predictive modelling workflow that can be improved by TREEOME integration. Position weight matrices can be combined with an epigenetic prior (*e.g.* H3K4me3 or DNase I hypersensitivity data) to identify putative transcription factor binding sites *in silico* using the Bayesian approach developed by [27]. This artificial data can be used to train a model in the same way as actual ChIP-seq data (see Section 2.2) to yield models of near-equivalent prediction accuracy [5]. The putative regulatory role of the transcription factor can then be derived using PCA, as described in Section 2.5.